2

AD

AD-E402 0*4

Technical Report ARFSD-TR-89014

A NEURAL NETWORK BASED SPEECH RECOGNITION SYSTEM

AD-A219 794

Edward J. Carroll
Norman P. Coleman, Jr.
G. N. Reddy

February 1990

US ARMY
ARMAMENT MUNITIONS
& CHEMICAL COMMAND
ARMAMENT RDE CENTER

U.S. ARMY ARMAMENT RESEARCH, DEVELOPMENT AND ENGINEERING CENTER

Fire Support Armaments Center

Picatinny Arsenal, New Jersey

Approved for public release; distribution unlimited.

90 08 2[illegible] 133

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER Technical Report ARFSD-TR-89014	5. MONITORING ORGANIZATION REPORT NUMBER
6a. NAME OF PERFORMING ORGANIZATION ARDEC, FSAC 6b. OFFICE SYMBOL SMCAR-FSF-RC	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (CITY, STATE, AND ZIP CODE) Fire Control Div Picatinny Arsenal, NJ 07806-5000	7b. ADDRESS (CITY, STATE, AND ZIP CODE)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION ARDEC, IMD STINFO Br 8b. OFFICE SYMBOL SMCAR-IMI-I	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (CITY, STATE, AND ZIP CODE) Picatinny Arsenal, NJ 07806-5000	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (INCLUDE SECURITY CLASSIFICATION)

A NEURAL NETWORK BASED SPEECH RECOGNITION SYSTEM

12. PERSONAL AUTHOR(S) Edward J. Carrol, Norman P. Coleman, Jr., and G. N. Reddy

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (YEAR, MONTH, DAY)	15. PAGE COUNT
Final	FROM Jun 89 TO Sep 89	February 1990	14

16 SUPPLEMENTARY NOTATION

Work completed in cooperation with G. N. Reddy, Michigan Technological University, Department of Electrical Engineering, Houghton, Michigan 49931.

17. COSATI CODES			18. SUBJECT TERMS (CONTINUE ON REVERSE IF NECESSARY AND IDENTIFY BY BLOCK NUMBER)
FIELD	GROUP	SUB-GROUP	Neural network Back-propagation Speech recognition

19. ABSTRACT (CONTINUE ON REVERSE IF NECESSARY AND IDENTIFY BY BLOCK NUMBER)

This report presents an overview of the development of a neural network based speech recognition system. The two primary tasks involved were the development of a time invariant speech encoder and a pattern recognizer or detector. The speech encoder uses amplitude normalization and a Fast Fourier Transform to eliminate amplitude and frequency shifts of acoustic clues. The detector consists of a back-propagation network which accepts data from the encoder and identifies individual words. This use of neural networks offers two advantages over conventional algorithmic detectors: the detection time is no more than a few network time constants, and its recognition speed is independent of the number of the words in the vocabulary. The completed system has functioned as expected with high tolerance to input variation and with error rates comparable to a commercial system when used in a noisy environment.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL I. HAZNEDARI	22b. TELEPHONE (INCLUDE AREA CODE) AV 880-3316 22c. OFFICE SYMBOL SMCAR-IMI-I

DD FORM 1473, 84 MAR

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

CONTENTS

INTRODUCTION

In spite of the tremendous progress in computer technology, speech recognition remains one of the most difficult task for computers to accomplish. Various factors contribute to this complexity which include speaker variability, large data samples, and excessive computational overhead in detection. Most of the speech recognition systems primarily consist of two major phases: time-invariant speech feature extraction and detection.

Variability from speaker to speaker is due to variations in loudness, rate, and dialect. The loudness problem can usually be taken care of with amplitude normalization. Solutions to the problem of rate or time have been attempted using time normalization (time-warping) (refs 1 and 2). The variability of speech due to dialect is very difficult to manage. Various measures have been attempted to alleviate this problem such as using average or multiple templates as a reference instead of a single template. Averaging the templates results in extraction of primary features of the speech signal. A detector which uses averaged templates is therefore less sensitive to minor local variations due to dialect. The recognition system then responds only to principal features of the whole word and becomes immune to local micro changes. Therefore, the system becomes more robust. Most of the averaging procedures used in earlier studies are time-domain techniques. Averaging in time is quite difficult due to problems in exact identification of the end points. Using multiple templates eliminates the need for averaging but slows down the conventional speech recognizer due to the increased number of patterns the system must examine before determining which word was spoken.

The second problem is the large amount of data needed for processing. The spectral range of speech lies approximately between 60 Hz and 4 kHz. With a sampling frequency of 8 kHz and word lengths as long as 1 sec, the data sample will contain 8,000 points. Substantial data reductions have been achieved through Linear Predictive Coding (LPC) (ref 3) and Short-term Fourier Analysis (ref 4). Further data reduction has been achieved through vectorization by replacing vectors with simple indexes (refs 4 and 5).

Finally, detection or finding the distance between the reference template and the input has been one of the major problems in speech recognition due to excessive computational overhead, especially for systems with a large vocabulary. The following section describes a typical conventional speech recognition system.

CONVENTIONAL SPEECH RECOGNITION SYSTEMS

A block diagram of a typical conventional speech recognition system is shown in figure 1. This involves three basic steps:

1. Speech coding or feature extraction: Computes spectral coefficients every 10 or so milliseconds using either a Fast Fourier Transform (FFT) or Linear Predictive Coding (LPC).

2. Time-Warping: Computes local frame-to-frame distances and uses these local distances to time align input sequences.

3. Detection: Computes the whole word matching score using the local distances to the corresponding reference work templates.

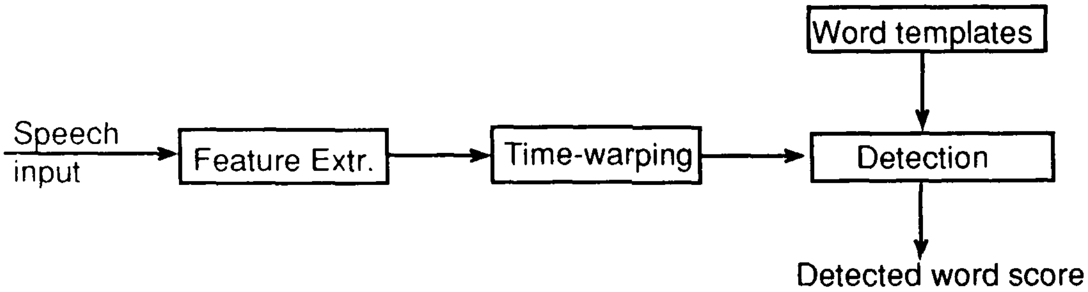

Figure 1. Three basic steps involved in conventional speech recognition

The problem with this system is the assumption that acoustic clues, in a given word/signal, appear in a precise time sequence; this is an erroneous assumption (ref 6). For a robust speech recognition system, it is essential to look for global word clues rather than local peaks and valleys. This is possible only if the analysis is based on entire word data rather than local data sets. A indicated earlier, the detector is computationally intensive. This computation is especially time consuming for large vocabulary systems since the input signal has to be compared with every reference template. The neural network based prototype system described here addresses both the problem of global feature extraction and the need for an improved detector to alleviate the problem of extensive computation.

NEURAL NETWORK BASED SPEECH RECOGNITION SYSTEM

A block diagram of the neural network based speech recognition system is shown in figure 2. The speech recognizer developed is neural network based.

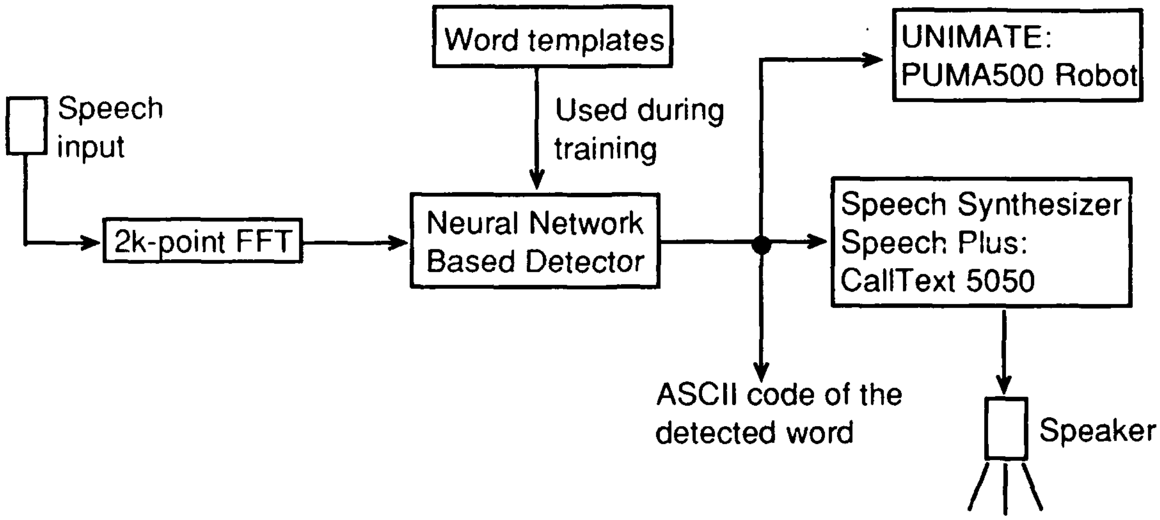

Figure 2. A neural network based speech recognition system

The system primarily consists of two sections: a time-invariant speech coder and a neural network based detector.

Time-Invariant Speech Coding

This section performs spectral feature extraction for a whole word. It is a 2k-point FFT program, giving a spectral range approximately between 2 Hz and 4 kHz. Each input was sampled at 8 kHz with a total sampling time of 0.50 sec. Most of the words are less than 0.50 sec duration. Whole word processing, unlike in conventional systems where a group of local data sets are used, eliminates the need for time-warping and end-point detection. The first half of the 2k-point FFT was averaged over each 8 consecutive points to give a compressed data of 256 points. This was needed to resolve some of the memory problems during computer simulation of the neural network detector.

Neural Network Detector

The detector was designed using a back-propagation (BP) neural network (refs 7 and 8). The network was trained for a set of 11 words used for controlling the UNIMATE PUMA-560 robot. The word set used was: Stop, Start, Exit, Forward, Backward, Right, Left, Up, Down, Open, and Close.

In normal test mode, the detector generates a coded equivalent of the word recognized. A program in the robot controller converts this coded sequence into an appropriate binary code for the robot to act. This coded sequence is then converted to an ASCII text string. It is also fed to a speech synthesizer (Speech Plus Inc.; Model: CallText 5050) for audio reproduction of the received commands.

The detector consists of a multilayer feed forward network with one input layer, two hidden layers, and one output layer. The network has 256 processing elements (PEs) in the input layer, 20 PEs in each of the hidden layers, and 11 processing elements in the output layer. The number of PEs in the input layer is equivalent to the number of outputs used from the FFT processor; the PEs in the hidden layers were chosen as a compromise between speed of training and representational power. The output layer consists of 11 PEs corresponding to the number of bits required to represent the maximum number of words in the test vocabulary. Details of the NN-based detector are included in the following section.

Back-Propagation Network Topology

A topological description of the NN-speech detector showing the number of layers used, number of PEs in each layer, the type of transfer functions used, and the learning rules for each layer is shown in figure 3. The description of the transfer functions and algorithms follow the topological description with the following used as the defaults:

Summation Fn (SF)	= None	Learning Rule (LR)	= Cumulative Delta
Scale	= 1.0	High Limit (HL)	= +1
Offset	= 0.0	Low Limit (LL)	= -1
Output Fn (OF)	= Direct	Transfer Fn (TF)	= Sigmoid Fn
Bias (B)	= 1		

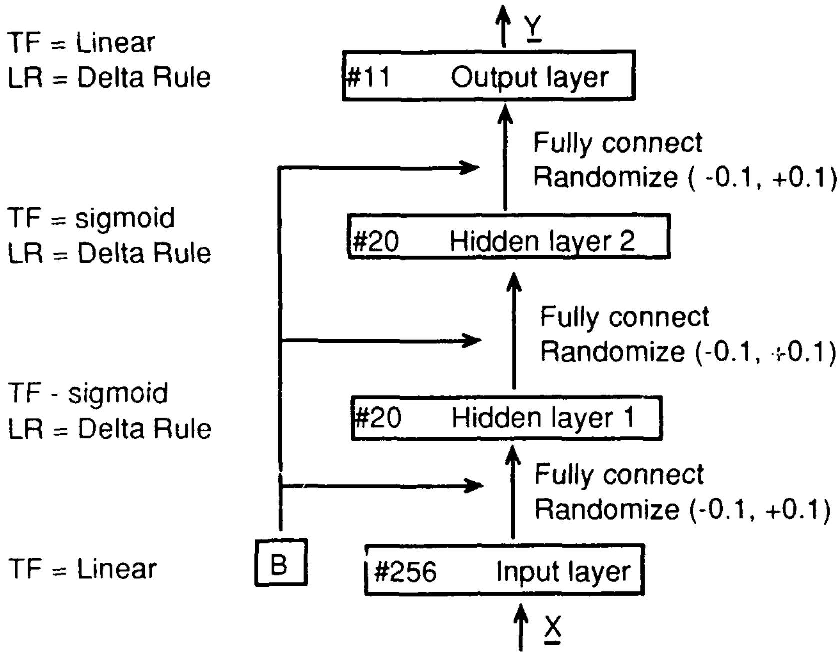

Figure 3. Back-propagation neural network topology

Transfer Functions

o Sigmoid

$$y = (1 + e^{-I * Gain})^{-1}$$

o Linear

$$y = I$$

where

I is the weighted sum of the inputs and is given by $I_j = \sum X_i * W_{ji}$

Training Algorithm

Cumulative-Delta Rule: All patterns are concurrently presented to the network. Weights are updated after each pattern is applied and the error corresponding to that pattern is computed (Tssp). When the weights are updated for each pattern once it is called an epoch. The network must go through many epochs before it becomes trained. The error is computed as the sum of the squares of error of all the patterns (Tss). Training stops when this cumulative error is below a specified thresh-hold (TssTh). When Tss reaches indicated TssTh the network is considered trained. This algorithm is presented in a step-wise fashion as follows:

Cumulative Back-Propagation Algorithm

1. Randomize all weights -0.1 and +0.1. Set number of patterns (P) and cumulative error to zero (P = 0, Tss = 0).

2. For a given input x_p, computed output y_p, and desired output d_p, apply input x_p and compute y_p.

3. Based on (d - y) update weights, compute y, Tssp, Tss.

Compute error at the output layer as:

δ at the output layer = δ^{s+1} = (d - y) y (1 - y); for TF = Sigmoid

$= (d - y) * \partial y . \partial I$

Compute errors for the hidden layers as:

δ for the hidden units = $\delta_k^s = y_k (1 - y_k) \sum_k \delta_k^{s+1} * W_k^s$

and update weights using:

$W_{ij}^s = W_{ij}^s + \varepsilon * \delta^{s+1} * X_{ij}^s$

Use graded training coefficient (ε ->1.0 to 0.1) depending on Tss (10->0.1).

Compute Tssp as: Tssp = $\sum(di - yi)^2 i = 1,2,...N$; where N is the number elements in the output layer and Tss as: Tss = Tss + Tssp

4. Set P = last pattern number trained

5. If Tss>TssTh for any pattern go to step 2; else END.

More details on BP neural network can be found in references 7 and 8.

TEST RESULTS

The speech recognition system was integrated into the PUMA-560 robot control loop; testing of the network, however, was done stand-alone in a noisy (electrical and acoustal) environment. As expected, it was capable of learning all patterns. Operating in speaker-dependent mode, it performed with an error rate of less than 12.7%. This high value, however, compares very favorably with an excellent conventional speech recognition system called VocalLink from Interstate Voice Products. This commercial system, tested under the same high noise conditions, exhibited an error rate of 14.5%.

CONCLUSIONS

A speech recognition system was developed by using a neural network for a detection stage. This stage, when implemented in hardware, will provide almost instant detection irrespective of the number of templates. This is contrary to the conventional techniques where detection is very computationally intensive especially with a large vocabulary. This research also addressed another major concern in speech recognition systems namely that of time-warping. Global Fast Fourier Transform transformations of the whole word provide automatic time warping and seems to perform better than conventional time-warping techniques and also eliminates the problem of end-point detection. Currently, the system is computer simulated; for real-time application the speech-coding part has to be hard-wired. The concept, however, works well.

RECOMMENDATIONS

To make the neural network detection more robust, it is preferable to develop the set of training templates at different times. The system also exhibits superior robustness to speech variations when training is done with individual words from a training sample set instead of single averaged word templates. Additional methods of speech detection and preprocessing should be examined with the network performing the detector fusion

REFERENCES

1. Myers, C.; Rabiner, R.; and Rosenberg, E., "Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-28, no. 6, pp 623-635, December 1980.

2. Hiroaki, S. and Chiba, S., "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-26, no. 1, pp 43-49, February 1978.

3. Markel, J. D. and Gray, A. H., Linear Predication of Speech, Springer-Verlag, New York, NY, 1976.

4. Lippmann, R. P., "Review of Neural Networks for Speech Recognition," Neural Computation, vol. 1, no. 1, pp 1-38, Spring 1989.

5. Duda, R. O. and Hart, P. E., Pattern Classification and Scene Analysis, John-Wiley & Sons, New York, NY, 1973.

6. Waibel, A., "Modular Construction of Time-Delay Neural Networks for Speech Recognition," Neural Computation, vol. 1, no. 1, pp 39-46, Spring 1989.

7. Wasserman, P. D., Neural Computing: Theory and Practive, Van Nostrand Reinhold, New York, NY, pp 43-60, 1989.

8. Neural Ware Inc. User's Guide: NeuralWorks Explorer, Sewickley, PA, pp 438-480, 1988.

9. Reddy, G. N.; Carroll, E. J.; and Coleman, N. P. Jr., "Applications of Neural Networks: A Review," International Conference on Expert Systems and Neural Networks, Honolulu, HI, August 16-18, 1989.

DISTRIBUTION LIST

Commander
Armament Research, Development and Engineering Center
U.S. Army Armament, Munitions and Chemical Command
ATTN: SMCAR-IMI-I (5)
SMCAR-FSF-RC
Picatinny Arsenal, NJ 07806-5000

Commander
U.S. Army Armament, Munitions and Chemical Command
ATTN: AMSMC-GCL(D)
Picatinny Arsenal, NJ 07806-5000

Administrator
Defense Technical Information Center
ATTN: Accessions Division (12)
Cameron Station
Alexandria, VA 22304-6145

Director
U.S. Army Materiel Systems Analysis Activity
ATTN: AMXSY-MP
Aberdeen Proving Ground, MD 21005-5066

Commander
Chemical Research, Development and Engineering Center
U.S. Army Armament, Munitions and Chemical Command
ATTN: SMCCR-MSI
Aberdeen Proving Ground, MD 21010-5423

Commander
Chemical Research, Development and Engineering Center
U.S. Army Armament, Munitions and Chemical Command
ATTN: SMCCR-RSP-A
Aberdeen Proving Ground, MD 21010-5423

Director
Ballistic Research Laboratory
ATTN: AMXBR-OD-ST
Aberdeen Proving Ground, MD 21005-5066

Chief
Benet Weapons Laboratory, CCAC
Armament Research, Development and Engineering Center
U.S. Army Armament, Munitions and Chemical Command
ATTN: SMCAR-CCB-TL
Watervliet, NY 12189-6000

Commander
U.S. Army Armament, Munitions and Chemical Command
ATTN: SMCAR-ESP-L
Rock Island, IL 61299-6000

Director
U.S. Army TRADOC Systems Analysis Activity
ATTN: ATAA-SL
White Sands Missile Range, NM 88002